

# Energy Efficient Design for Tactile Internet

Changyang She and Chenyang Yang

School of Electronics and Information Engineering, Beihang University, Beijing, China

Email: {cyshe,cyyang}@buaa.edu.cn

**Abstract**—Ensuring the ultra-low end-to-end latency and ultra-high reliability required by tactile internet is challenging. This is especially true when the stringent Quality-of-Service (QoS) requirement is expected to be satisfied not at the cost of significantly reducing spectral efficiency and energy efficiency (EE). In this paper, we study how to maximize the EE for tactile internet under the stringent QoS constraint, where both queueing delay and transmission delay are taken into account. We first validate that the upper bound of queueing delay violation probability derived from the effective bandwidth can be used to characterize the queueing delay violation probability in the short delay regime for Poisson arrival process. However, the upper bound is not tight for short delay, which leads to conservative designs and hence leads to wasting energy. To avoid this, we optimize resource allocation that depends on the queue state information and channel state information. Analytical results show that with a large number of transmit antennas the EE achieved by the proposed policy approaches to the EE limit achieved for infinite delay bound, which implies that the policy does not lead to any EE loss. Simulation and numerical results show that even for not-so-large number of antennas, the EE achieved by the proposed policy is still close to the EE limit.

## I. INTRODUCTION

Tactile internet enables unprecedented mobile applications such as vehicle collision avoidance, mobile robots, virtual reality and augmented reality [1, 2], which calls for ultra-low latency (say 1 ms) and ultra-high reliability (say 99.99999%). To ensure the low end-to-end (E2E) delay and high reliability for each short packet, both transmission and process delay and queueing delay should be bounded with small violation probability [3], and the delay spent in the backbone network should be controlled by updating the network architectures. By introducing short frame structure, short transmit time interval (TTI) [4] and using short codes such as Polar codes [5], the transmission, processing and coding delay can be reduced.

Though satisfying such a stringent quality of service (QoS) itself is rather challenging, it is not expected to be achieved at the cost of significantly reducing the spectral efficiency and energy efficiency (EE), which are important metrics for the fifth generation (5G) networks [6]. To guarantee such a stringent QoS, the resource allocation could be conservative, which may leads to a waste of energy. Moreover, to ensure the delay that may even be shorter than the channel coherence time, channel inversion power allocation is required in single-user case, which leads to unbounded transmit power. This suggests that the EE of tactile internet systems may be low. As far as the authors known, the EE related issues has not been considered in the context of in tactile internet.

The QoS requirement of tactile internet can be characterized by a delay bound (say, including air interface delay and

queue delay) and a delay bound violation probability (say, including the queueing delay violation probability, packet loss, and drop probability). Improving the EE under the queueing delay bound and delay bound violation probability constraint has been widely studied in existing studies, e.g., [7, 8]. Effective bandwidth and effective capacity is a powerful tool in designing resource allocation under such a statistical delay requirement [9]. However, since the distribution of queueing delay is obtained based on large deviation principle, effective bandwidth is widely believed useful only for optimizing the system with large delay requirement. It is unclear whether it can be used for design tactile internet with the short delay.

In this paper, we study how to maximize EE by optimizing resource allocation under the QoS provision of tactile internet. We validate that the effective bandwidth can be used as a tool in the short delay regime. In fact, for the applications with ultra-low latency, an upper bound of queueing delay violation probability derived from effective bandwidth can be applied for Poisson process and the arrival processes that are more bursty than Poisson [10]. However, the upper bound of the queueing delay violation probability is not tight, which inevitably leads to conservative design. To avoid wasting energy by the conservative designs, a queue state information (QSI) and channel state information (CSI) dependent resource allocation policy is proposed. Our analysis shows that the proposed policy is optimal in large scale antenna systems, and can achieve the EE limit obtained for the infinite delay bound. This implies that ensuring the ultra-low E2E delay and ultra-high reliability will to cause EE loss if the optimal policy is applied. As a by-product, we also derive the bandwidth and power required to guarantee the QoS. Simulation and numerical results validate our analysis and show that even with not-so-large number of antennas, the achieved EE of the proposed policy is closed to the EE limit.

## II. SYSTEM MODEL

Consider a time division duplexing cellular system, where  $K$  single-antenna users are served by a BS with  $N_t$  antennas during successive frames. Each frame is with duration  $T_f$ , which consists of a downlink (DL) and an uplink (UL) transmission phase. In the UL phase, each user (i.e., a vehicle) uploads its safety messages (e.g., speed and location [11]) with short packets to the BS. In the DL phase, the BS sorts the received safety messages from the nearby users of each user, and then transmits the relevant messages to the target users. To capture the essence of the problem, we consider

frequency division multiple access to avoid the interference among multiple users.

For the tactile service, the QoS can be characterized by an E2E delay bound for each packet,  $D_{\max}$ , and a packet loss/error probability,  $\varepsilon_D$ . The E2E delay is very short, say 1 ms [1], which includes UL and DL transmission delay, processing and coding delay, and queueing delay in the buffer of BS. To ensure the transmission delay ultra-low, we consider the short frame structure proposed in [4], where the TTI is the same as the frame duration and  $T_f \ll D_{\max}$ . Moreover, we assume that some sort of short codes can be applied such that the processing and coding delay is very low. Since the packet size is small (say less than 100 bytes), UL and DL transmission of each packet can be finished within a frame [4]. As a consequence, the maximal queueing delay of each packet allowed by the service is  $D_{\max}^q \triangleq D_{\max} - T_f$ . Denote the maximal queueing delay violation probability allowed by the service as  $\varepsilon^q$ . Then, the requirement imposed on the queueing delay for each packet is  $(D_{\max}^q, \varepsilon^q)$ , where  $\varepsilon^q < \varepsilon_D$ .

Consider block fading channel, which remains constant within each coherent interval of duration  $T_c$  and changes independently among the intervals. For the users with velocities of 30 ~ 120 km/h and the system operating in carrier frequency of 2 GHz, the channel coherence time is around 1.1 ~ 4.5 ms, which is larger than the queueing delay of each packet (i.e.,  $T_c > D_{\max}^q$ ), as illustrated in Fig. 1. In this paper, we consider such a typical scenario of vehicular communication systems, which is more challenging than the other case with  $T_c \leq D_{\max}^q$  in terms of stringent delay performance. For notational simplicity,  $T_c$  is assumed to be divisible by  $T_f$ .

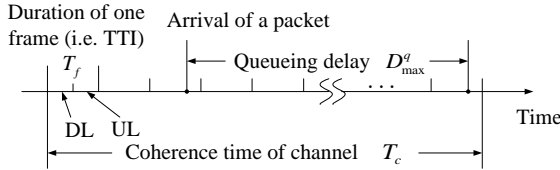


Fig. 1. Relation of delay bound and coherence time.

Due to the low rate requirement of each user, the bandwidth allocated to each user is usually less than the coherent bandwidth of the channel. Hence, it is reasonable to assume flat fading. Denote the average channel gain and channel vector of the  $k$ th user in a certain coherence interval as  $\alpha_k$  and  $\mathbf{h}_k \in \mathbb{C}^{N_t \times 1}$ , whose elements are independent and identically Gaussian distributed with zero mean and unit variance. When  $\alpha_k$  and  $\mathbf{h}_k$  are perfectly known at the BS and the user, the maximal number of packets that *can be* transmitted to the  $k$ th user in the  $n$ th frame is given by

$$s_k(n) = \frac{\Phi T_D W_k(n)}{u} \log_2 \left[ 1 + \frac{\alpha_k P_k^t(n) g_k}{N_0 W_k(n)} \right] \text{ (packets)}, \quad (1)$$

where  $u$  is the size of each packet,  $P_k^t(n)$  and  $W_k(n)$  are respectively the transmit power and bandwidth allocated to the  $k$ th user according to its queue length in the  $n$ th frame,  $T_D$  is the duration of DL transmission phase,  $N_0$  is the single-sided noise spectral density,  $g_k = \mathbf{h}_k^H \mathbf{h}_k$  is instantaneous channel gain,  $[\cdot]^H$  denotes the conjugate transpose, and  $\Phi \in (0, 1]$  is

the gap between channel capacity and data rate achieved by finite blocklength codes under given error probability  $\varepsilon^c$  [12].

In the  $n$ th frame, the  $k$ th user requests the packets uploaded from its nearby users, whose indices constitute a set  $\mathcal{A}_k$  with cardinality  $|\mathcal{A}_k|$ . As illustrated in Fig. 2, the index set of the nearby users of the  $k$ th user is  $\mathcal{A}_k = \{k+1, \dots, k+m\}$ . Then, the number of packets waited in the queue for the  $k$ th user at the beginning of the  $(n+1)$ th frame can be expressed as

$$Q_k(n+1) = \max \{Q_k(n) - s_k(n), 0\} + \sum_{i \in \mathcal{A}_k} a_i(n), \quad (2)$$

where  $a_i(n)$ ,  $i \in \mathcal{A}_k$  is the number of packets uploaded to the BS from the  $i$ th nearby user of the  $k$ th user.

We consider the scenario that the inter-arrival time between packets could be shorter than  $D_{\max}^q$  (otherwise the queueing delay is zero), which happens when the packets for a target user are randomly uploaded from multiple nearby users, i.e.,  $|\mathcal{A}_k| > 1$ . Intuitively, such a scenario seems to occur with a low probability. However, to ensure the ultra-high reliability of  $\varepsilon_D = 0.001\% \sim 0.00001\%$  [1, 2], the scenario of non-zero queueing delay is not negligible.

Denote the number of packets departed from the  $k$ th queue in the  $n$ th frame as  $b_k(n)$ . If all the packets in the queue can be successfully transmitted in the  $n$ th frame, then  $b_k(n) = Q_k(n)$ . Otherwise,  $b_k(n) = s_k(n)$ . Hence, we have

$$b_k(n) = \min \{Q_k(n), s_k(n)\}. \quad (3)$$

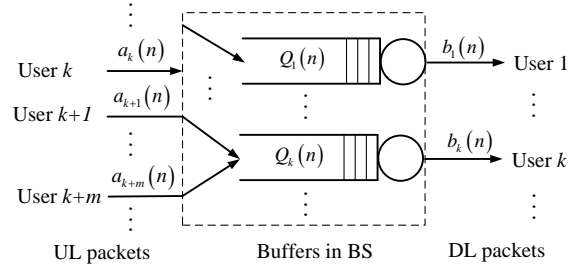


Fig. 2. Queueing model at the BS.

Considering (2) and (3), we can show after some regular derivations that the queue length evolves as follows

$$Q_k(n+1) - Q_k(n) = \sum_{i \in \mathcal{A}_k} a_i(n) - b_k(n), \quad (4)$$

which implies that the queueing delay can be controlled by adjusting the departure process.

### III. ENSURING THE QUEUEING DELAY REQUIREMENT

In this section, we employ *effective bandwidth* to represent the QoS constraint imposed on the queueing delay. Then, we present a M/D/1 queueing model, with which we validate that effective bandwidth can be applied in the short delay regime.

#### A. Representing QoS Constraint with Effective Bandwidth

The aggregation of the packet arrival processes from the  $|\mathcal{A}_k|$  nearby users of the  $k$ th user (i.e.,  $\sum_{i \in \mathcal{A}_k} a_i(n)$  in (2)) can

be modelled as a Poisson process [11]. For a Poisson arrival process, the effective bandwidth is [3]

$$E_k^B(\theta_k) = \frac{\lambda_k}{T_f \theta_k} (e^{\theta_k} - 1) \text{ (packets/s)}, \quad (5)$$

where  $\theta_k$  is the QoS exponent,  $\lambda_k$  is the average number of packets arrived at the  $k$ th queue during one frame, which is identical for all frames.

When the  $k$ th user is served with a constant rate equal to  $E_k^B(\theta_k)$ , the steady state queueing delay violation probability can be approximated as [13]

$$\Pr\{D_k(\infty) > D_{\max}^q\} \approx \eta_k \exp\{-\theta_k E_k^B(\theta_k) D_{\max}^q\}, \quad (6)$$

where  $\eta_k$  is the buffer non-empty probability, and the approximation is accurate when  $D_{\max}^q \rightarrow \infty$  [9].

Since  $\eta_k \leq 1$ , we have

$$\Pr\{D_k(\infty) > D_{\max}^q\} \leq \exp\{-\theta_k E_k^B(\theta_k) D_{\max}^q\} \triangleq P_{D_k}^{\text{UB}}. \quad (7)$$

If the upper bound in (7) satisfies

$$P_{D_k}^{\text{UB}} = \exp\{-\theta_k E_k^B(\theta_k) D_{\max}^q\} = \varepsilon^q, \quad (8)$$

then the QoS requirement  $(D_{\max}^q, \varepsilon^q)$  can be satisfied. We can obtain  $\theta_k$  from (8) for a service with given QoS requirement and effective bandwidth, which is a key parameter in the QoS constraint imposed on resource allocation.

Substituting (5) into (8), we can obtain that  $\theta_k = \ln \left[ \frac{T_f \ln(1/\varepsilon^q)}{\lambda_k D_{\max}^q} + 1 \right]$ . With which (5) can be re-expressed as,

$$E_k^B(\theta_k) = \frac{\ln(1/\varepsilon^q)}{D_{\max}^q \ln \left[ \frac{T_f \ln(1/\varepsilon^q)}{\lambda_k D_{\max}^q} + 1 \right]} \text{ (packets/s)}. \quad (9)$$

To guarantee  $(D_{\max}^q, \varepsilon^q)$ , the minimal number of packets transmitted to the  $k$ th user in the  $n$ th frame should be a constant among frames that satisfies [9]

$$s_k(n) = T_f E_k^B(\theta_k) \text{ (packets)}. \quad (10)$$

When the  $k$ th queue is served by the constant service process  $\{s_k(n), n = 1, 2, \dots\}$  that satisfies (10), the departure process in (3) becomes

$$b_k(n) = \min\{Q_k(n), T_f E_k^B(\theta_k)\} \text{ (packets)}. \quad (11)$$

Therefore, if the departure process  $\{b_k(n), n = 1, 2, \dots\}$  satisfies (11), then  $(D_{\max}^q, \varepsilon^q)$  can be guaranteed.

#### B. Validating the Upper Bound $P_{D_k}^{\text{UB}}$ with M/D/1 Model

As defined in [9], the effective bandwidth is applicable for the scenario when  $D_{\max}^q \rightarrow \infty$ . In other words, the approximation in (6) is accurate when the delay bound is sufficient large. However, it is unclear when the value of  $D_{\max}^q$  is large enough for an accurate approximation. One possible reason is that it is very difficult to obtain an accurate distribution of the queue length or queueing delay.

Yet the real concern for the problem at-hand is whether the upper bound in (7) is applicable. If  $P_{D_k}^{\text{UB}}$  is indeed an upper bound of  $\Pr\{D_k(\infty) > D_{\max}^q\}$ , then a transmit policy

satisfying the QoS constraint in (10) or (11) can guarantee the required QoS.

When a Poisson arrival process is served by a constant service process  $\{s_k(n), n = 1, 2, \dots\}$ , the well-known M/D/1 queueing model can be applied [14]. For a discrete state M/D/1 queue with length as integer (i.e., the number of packets), the closed-form expression of the queue length distribution is known. Specifically, the complimentary cumulative distribution function (CCDF) of the steady state queue length can be expressed as  $\Pr\{Q_k(\infty) > l\} = 1 - \sum_{i=1}^l \pi_i$ , where  $\pi_l = \Pr\{Q_k(\infty) = l\}$  is the probability that there are  $l$  packets in the queue, which is [14],

$$\begin{aligned} \pi_0 &= 1 - \gamma_k, \quad \pi_1 = (1 - \gamma_k)(e^{\gamma_k} - 1), \\ \pi_l &= (1 - \gamma_k) \times \\ &\left\{ e^{l\gamma_k} + \sum_{i=1}^{l-1} e^{i\gamma_k} (-1)^{l-i} \left[ \frac{(i\gamma_k)^{l-i}}{(l-i)!} + \frac{(i\gamma_k)^{l-i-1}}{(l-i-1)!} \right] \right\}, \\ &(l \geq 2), \end{aligned} \quad (12)$$

with  $\gamma_k = \lambda_k/s_k(n)$ . For a Poisson arrival process served by a constant service process, the CCDF of the queueing delay can be derived from Appendix D in [8] as

$$\Pr\{D_k(\infty) > T_f l/s_k(n)\} = 1 - \sum_{i=1}^l \pi_i. \quad (13)$$

To derive a QoS constraint for resource allocation, we need to derive the expression of  $s_k(n)$  as a function of  $D_{\max}^q$  and  $\varepsilon^q$  by setting  $l = s_k(n) D_{\max}^q / T_f$  and  $1 - \sum_{i=1}^l \pi_i = \varepsilon^q$ . However, the expression of  $\pi_l$  in (12) is complex. Thus, the expression of  $s_k(n)$  cannot be obtained in closed-form. This indicates that the M/D/1 model is hard to be used for optimizing a transmit policy to ensure the QoS. Nonetheless, (13) can be used to validate the upper bound  $P_{D_k}^{\text{UB}}$  in (7) via numerical results.

#### IV. ENERGY-EFFICIENT RESOURCE ALLOCATION

The EE is the ratio of the amount of successfully transmitted data to the energy consumption [15], i.e.,  $EE = (1 - \varepsilon_D) \sum_{k=1}^K \sum_{i \in \mathcal{A}_k} \mathbb{E}[a_i(n)] / \{T_D \mathbb{E}[P_{\text{tot}}(n)]\}$ , where  $P_{\text{tot}}(n)$  is the total power consumed at a BS for DL transmission in the  $n$ th frame, which can be modeled as [16]

$$P_{\text{tot}}(n) = \frac{1}{\rho} \sum_{k=1}^K P_k^t(n) + P^{cw} \sum_{k=1}^K W_k(n) + P_0^c, \quad (14)$$

where  $\rho \in (0, 1]$  is the power amplifier efficiency,  $P^{cw}$  is the circuit power consumption per unit bandwidth, and  $P_0^c$  is the circuit power that is independent of bandwidth.

Since the value of  $\varepsilon_D$  is very low, the nominator is almost independent of transmit policy. Hence, maximizing the EE is equivalent to minimizing the average total power consumption. To this end, we can minimize the instantaneous power consumption  $P_{\text{tot}}(n)$  by optimizing resource allocation.

### A. Queue Length Dependent Resource Allocation

Recall that the service process  $\{s_k(n), n = 1, 2, \dots\}$  should be a constant among successive frames satisfying (10), in order to ensure the queueing delay requirement. To support such a constant service process, it is shown from (11) that when  $Q_k(n) < T_f E_k^B(\theta_k)$ ,  $b_k(n) < s_k(n)$ . This indicates that the number of departed packets may be less than the number of packets that *can be* transmitted by the system in some frames. To save energy, i.e., avoid wasting resources of the system, we introduce a queue length dependent two-state policy: when  $Q_k(n) > T_f E_k^B(\theta_k)$ ,  $s_k(n) = T_f E_k^B(\theta_k)$ , otherwise  $s_k(n) = Q_k(n)$ , i.e.,

$$s_k(n) = \min\{Q_k(n), T_f E_k^B(\theta_k)\}. \quad (15)$$

Substituting (15) into (3), we can show that the departure process has the same form as in (11). This means that if a two-state policy satisfies (15), then  $(D_{\max}^q, \varepsilon^q)$  can be guaranteed.

From (1), (15) and (14), the two-state transmit power and bandwidth allocation policy that minimizes the instantaneous total power consumption under the constraint imposed on  $(D_{\max}^q, \varepsilon^q)$  can be obtained from the following problem,

$$\begin{aligned} & \min_{\substack{P_k^t(n), W_k(n), \\ k=1, \dots, K}} \sum_{k=1}^K P_k^t(n) + P^{cw} \rho \sum_{k=1}^K W_k(n) \quad (16) \\ & \text{s.t. } \frac{\Phi T_D W_k(n)}{u} \log_2 \left[ 1 + \frac{\alpha_k P_k^t(n) g_k}{N_0 W_k(n)} \right] \\ & \quad = \min\{Q_k(n), T_f E_k^B(\theta_k)\}, \quad k = 1, \dots, K. \quad (16a) \end{aligned}$$

To show how much resource is required to guarantee the stringent QoS, the maximal transmit power and bandwidth constraints are not considered.

To solve the problem, we relax (16a) into inequality constraints, and refer to the new problem that minimizes (16) under the inequality constraints as *Problem A*. It is not hard to show that *Problem A* is equivalent to the original problem. Because the left hand side of (16a) is jointly concave in  $P_k^t(n)$  and  $W_k(n)$ , *Problem A* is convex, which can be solved by standard tools such as the interior-point method [17].

### B. Optimality of the Two-state Policy and Required Resources

The two-state policy is heuristic, since a policy with more than two states may give rise to lower power consumption. Nonetheless, in the sequel we show that the optimized two-state policy can maximize the EE in a large  $N_t$  asymptotic. With the resulting closed-form solution, we can show how much resources are required to ensure the QoS in such a region. The simulations later show that the results obtained for large value of  $N_t$  also hold when  $N_t$  is not so large.

1) *Minimal Average Total Power Consumed by the Two-state Policy:* When  $N_t \rightarrow \infty$ , (1) approaches [18]

$$s_k(n) = \frac{\Phi T_D W_k(n)}{u} \log_2 \left[ 1 + \frac{\alpha_k N_t P_k^t(n)}{N_0 W_k(n)} \right] \text{ (packets).}$$

Due to channel hardening, the small scale channel fading does not affect the service process. In this case, the QoS constraint

can be obtained from (16a) by replacing  $g_k$  with  $N_t$ . The total power minimization resource allocation problem under such a QoS constraint is refer to as *Problem B*.

By analyzing the Karush-Kuhn-Tucker (KKT) conditions of *Problem B* using a similar way as the proof of Proposition 2 in [19], we can derive that the ratio of the optimal transmit power to the optimal bandwidth allocated to each user is a constant depending on  $N_t$ ,  $P^{cw}$ ,  $\rho$  and  $\alpha_k$ , i.e.,  $\frac{P_k^{t*}(n)}{W_k^*(n)} = P_k^{tw}$ . Then, we can find the optimal solution of *Problem B* as follows,

$$P_k^{t*}(n) = \frac{P_k^{tw} u \min\{Q_k(n), T_f E_k^B(\theta_k)\}}{\Phi T_D \log_2 \left( 1 + \frac{\alpha_k N_t P_k^{tw}}{N_0} \right)}, \quad (17)$$

$$W_k^*(n) = \frac{u \min\{Q_k(n), T_f E_k^B(\theta_k)\}}{\Phi T_D \log_2 \left( 1 + \frac{\alpha_k N_t P_k^{tw}}{N_0} \right)}. \quad (18)$$

Substituting  $P_k^{t*}(n)$  and  $W_k^*(n)$  into (14), we can obtain the minimal total power consumed by the two-state policy as

$$P_{\text{tot}}^*(n) = \sum_{k=1}^K \frac{\left( \frac{P_k^{tw}}{\rho} + P^{cw} \right) u \min\{Q_k(n), T_f E_k^B(\theta_k)\}}{\Phi T_D \log_2 \left( 1 + \frac{\alpha_k N_t P_k^{tw}}{N_0} \right)} + P_0^c.$$

With the two-state policy, (11) can be satisfied and hence  $\mathbb{E}[b_k(n)] = \mathbb{E}\{\min\{Q_k(n), T_f E_k^B(\theta_k)\}\}$ . Moreover, with the ensured QoS, for ergodic arrival and departure processes we have  $\mathbb{E}[b_k(n)] = (1 - \varepsilon_D) \sum_{i \in \mathcal{A}_k} \mathbb{E}[a_i(n)]$ . Then,  $\mathbb{E}[P_{\text{tot}}^*(n)]$  consumed by the optimal two-state policy can be rewritten as

$$\sum_{k=1}^K \frac{\left( \frac{P_k^{tw}}{\rho} + P^{cw} \right) u (1 - \varepsilon_D) \sum_{i \in \mathcal{A}_k} \mathbb{E}[a_i(n)]}{\Phi T_D \log_2 \left( 1 + \frac{\alpha_k N_t P_k^{tw}}{N_0} \right)} + P_0^c. \quad (19)$$

2) *A Lower Bound of Power Consumption:* To show that the optimized two-state policy is EE-optimal, we compare  $\mathbb{E}[P_{\text{tot}}^*(n)]$  with a lower bound achieved when  $D_{\max}^q \rightarrow \infty$ . The minimal average power consumption obtained for  $D_{\max}^q \rightarrow \infty$  is the ultimate lower bound of those for arbitrary finite  $D_{\max}^q$  requirements, and the resulting EE is the EE limit.

As shown in [19], to ensure the QoS with infinite delay bound, the service process only needs to satisfy

$$s_k(n) = (1 - \varepsilon_D) \sum_{i \in \mathcal{A}_k} \mathbb{E}[a_i(n)]. \quad (20)$$

The lower bound can be obtained by minimizing (16) under constraint (20). We refer to this problem as *Problem C*. By analyzing the KKT conditions of *Problem C* using a similar way as the proof of Proposition 2 in [19], we can show that the lower bound is the same as (19).

3) *Required Maximal Transmit Power and Bandwidth:* With the closed form solution of the optimal two-state policy, we can find the required resources to maximize the EE with guaranteed  $(D_{\max}^q, \varepsilon^q)$ , in order to provide guidance for designing systems serving tactile internet. The maximal transmit power and bandwidth to achieve the EE limit with

ensured QoS can be obtained respectively from,

$$P_{\text{req}}^t = \max_{n=1,2,\dots} \sum_{k=1}^K P_k^{t*}(n), W_{\text{req}} = \max_{n=1,2,\dots} \sum_{k=1}^K W_k^*(n). \quad (21)$$

Substituting (17) and (18) into (21), further considering  $\min\{Q_k(n), T_f E_k^B(\theta_k)\} \leq T_f E_k^B(\theta_k)$  and the expression of  $E_k^B(\theta_k)$  in (9), we can derive that

$$P_{\text{req}}^t \leq \sum_{k=1}^K \frac{P_k^{tw} u T_f [\ln(1/\varepsilon^q)] / (\Phi T_D D_{\text{max}}^q)}{\log_2 \left( 1 + \frac{\alpha_k N_t}{N_0} P_k^{tw} \right) \ln \left[ \frac{T_f \ln(1/\varepsilon^q)}{\lambda_k D_{\text{max}}^q} + 1 \right]}, \quad (22)$$

$$W_{\text{req}} \leq \sum_{k=1}^K \frac{u T_f [\ln(1/\varepsilon^q)] / (\Phi T_D D_{\text{max}}^q)}{\log_2 \left( 1 + \frac{\alpha_k N_t}{N_0} P_k^{tw} \right) \ln \left[ \frac{T_f \ln(1/\varepsilon^q)}{\lambda_k D_{\text{max}}^q} + 1 \right]}, \quad (23)$$

which are nearly proportional to  $1/D_{\text{max}}^q$  and  $\ln(1/\varepsilon^q)$ .

## V. SIMULATION AND NUMERICAL RESULTS

In this section, we first validate our analysis, and then show the resources required to guarantee the QoS of tactile internet with simulation and numerical results.

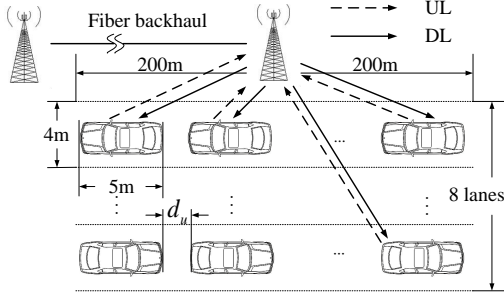


Fig. 3. Simulation scenario.

We consider an eight-lane two-direction highway scenario in urban area. The users (i.e., vehicles) uniformly located in the eight lanes are served by the roadside BSs with distance 400 meters who are connected by fiber backhaul. The packet delay caused by fiber backhaul is around  $D_B = 0.1$  ms [20]. The path loss model is  $10 \log_{10} \alpha_k = 35.3 + 37.6 \log_{10} d_k$ , where  $d_k$  is the distance between a BS and the  $k$ th user in meters. Each vehicle requests safe messages from other vehicles with distances less than 100 m. For the vehicles in the cell edge who request the messages from the vehicles in adjacent cells, the BS in the adjacent cell forwards the received messages to the BS who serves the user requesting the messages, then  $D_B$  is also counted in the E2E delay, i.e.,  $D_{\text{max}}^q + D_B + T_f \leq D_{\text{max}}$ . Since there are other factors except the queueing delay violation lead to packet loss and error (e.g., finite blocklength channel coding), here we set  $\varepsilon^q = \varepsilon_D/2$  for simplicity. Parameters in the sequel are listed in Table I, unless otherwise specified.

The CCDF of queueing delay for the packets to the  $k$ th user is shown in Fig. 4. The upper bound in (7), i.e.,  $\Pr\{D_k(\infty) > D_{\text{th}}\} \leq \exp\{-\theta_k E_k^B(\theta_k) D_{\text{th}}\}$ , is numerically obtained with different values of  $D_{\text{th}} \in [0, D_{\text{max}}^q]$ . The CCDF of delay with the discrete state M/D/1 model is numerically

TABLE I  
LIST OF SIMULATION PARAMETERS [1, 12, 16]

End-to-end delay $D_{\text{max}}$	1 ms
Reliability $1 - \varepsilon_D$	99.99999%
Frame duration $T_f$	0.1 ms
Duration of DL phase $T_D$	0.05 ms
Coherence time of channel $T_c$	2 ms
Packet size $u$	20 bytes
UL average packet arrival rate	20 packets/s/user
Data rate gap $\Phi$	0.9
Single-sided noise spectral density $N_0$	-173 dBm/Hz
Circuit power per unit bandwidth $P^{cw}$	$72N_t$ mW/MHz
Other circuit power (e.g., cooling) $P_0^c$	$136N_t$ mW

obtained from (13) with  $s_k(n) = E_k^B(\theta_k)$ . The simulation results are obtained by computing the queueing delay of the packets served by the optimal two-state policy (the solution of problem (16)) during  $10^9$  frames. It is shown that the simulated CCDFs are not smooth for short delay bounds, since the approximation in (6) is not accurate. However, the upper bound in (7) always exceeds the CCDFs obtained with the M/D/1 model and the simulated CCDFs, which indicates that  $P_{D_k}^{\text{UB}}$  is indeed an upper bound of queueing delay violation probability even when the delay bound is very short. This can be explained as follows. As shown in (9),  $E_k^B(\theta_k)$  increases with  $1/D_{\text{max}}^q$ . With a policy that ensures  $s_k(n) = E_k^B(\theta_k)$ ,  $s_k(n)$  also increases with  $1/D_{\text{max}}^q$ . As shown in (12),  $\pi_0$  increases with  $s_k(n)$ . As a result,  $\eta_k = 1 - \pi_0$  decreases with  $s_k(n)$  and hence increases with  $D_{\text{max}}^q$ . When  $D_{\text{max}}^q$  is small,  $\eta_k \ll 1$  (around 0.1 in the scenario of Fig. 4), which leads to a loose upper bound. This indicates that the QoS constraint derived from the upper bound is conservative.

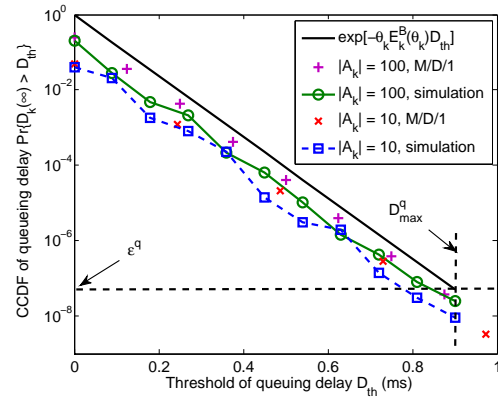


Fig. 4. Validating the upper bound in (7),  $N_t = 8$ .

TABLE II  
AVERAGE POWER CONSUMPTION WITH FINITE  $N_t$

$N_t$	2	4	8	16	32
Normalized $\mathbb{E}[P_{\text{tot}}(n)]$	1.149	1.042	1.015	1.005	1.002
Normalized $P_{\text{req}}^t$	0.983	0.482	0.458	0.442	0.436
Normalized $W_{\text{req}}$	0.463	0.424	0.419	0.414	0.412

To validate that the results obtained for large value of  $N_t$  are also true for not-so-large  $N_t$ , in Table II we provide the

simulation results of the average total power consumption, required transmit power and required bandwidth with finite  $N_t$ , normalized by those obtained with  $N_t \rightarrow \infty$  in (19), (22) and (23), respectively. To obtain the results, we solve problem (16) in  $2 \times 10^6$  frames (i.e.,  $10^5$  channel fading blocks) and then compute the averaged total power consumption, the maximal required transmit power and bandwidth. We set  $d_u = 15$  m,  $K = 160$  and  $|\mathcal{A}_k| = 80$  for the simulation.

The results in Table II show that the average power consumption when  $N_t > 2$  is close to the lower bound in (19). This indicates that the two-state policy is nearly EE-optimal, despite that the introduced QoS constraint is conservative. We can observe that the required maximal transmit power and bandwidth are less than the upper bounds in (22) and (23). This is because the upper bounds are obtained under the assumption that all the buffers are not empty. However, as show in Fig. 4, there is very high probability that a buffer is empty. When  $K$  is large, the number of users that have non-empty buffers is much less than  $K$ . Therefore, the required total transmit power and bandwidth is less than the upper bound.

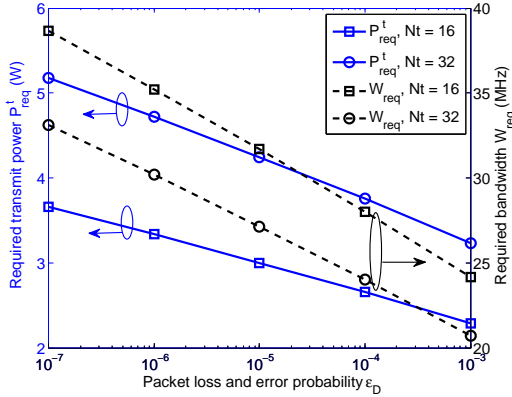


Fig. 5. Required maximal transmit power and bandwidth vs. reliability,  $D_{\max} = 1$  ms,  $d_u = 15$  m,  $K = 160$  and  $|\mathcal{A}_k| = 80$ .

In Fig. 5, we provide the upper bounds of the maximal transmit power and bandwidth required to achieve the EE limit with guaranteed QoS, which are numerically obtained from the right-hand side of (22) and (23). The results show that the required resources linearly increases with  $\ln(1/\epsilon_D)$ . This means that approaching the EE limit under the ultra-high reliability and ultra-low latency requirement does not need high transmit power or large bandwidth. We can see that the required bandwidth decreases with  $N_t$ , but the required transmit power increases with  $N_t$ . This can be explained as follows. Since  $P^{cw}$  increases with  $N_t$ , less bandwidth should be used to reduce circuit power consumption when  $N_t$  is large. With less bandwidth, more transmit power is required to ensure QoS.

## VI. CONCLUSION

In this paper, we studied how to design energy efficient resource allocation in tactile internet. To ensure the delay bound and its violation probability, an upper bound of the CCDF of queueing delay derived based on the effective bandwidth was applied. We optimized a QSI and CSI dependent

resource allocation policy to maximize the EE under the QoS constraint. We then showed that the minimal average total power consumption achieved by the optimized policy under the strict delay requirement equals to that under the infinite queueing delay requirement with large number of transmit antennas, which implies that the policy is optimal in maximizing the EE. Simulation and numerical results validated our analysis and showed that the achieved EE of the proposed resource allocation policy is closed to the upper bound of EE even for small number of antennas.

## REFERENCES

- [1] G. P. Fettweis, "The tactile internet: Applications & challenges," *IEEE Vehic. Tech. Mag.*, vol. 9, no. 1, pp. 64 – 70, Mar. 2014.
- [2] A. Osseiran, F. Boccardi and V. Braun, *et al.*, "Scenarios for 5G mobile and wireless communications: The vision of the METIS project," *IEEE Commun. Mag.*, vol. 52, no. 5, pp. 26 – 35, May. 2014.
- [3] C. She and C. Yang, "Ensuring the quality-of-service of tactile internet," in *Proc. IEEE VTC Spring*, 2016.
- [4] P. Kela and J. Turkka, *et al.*, "A novel radio frame structure for 5G dense outdoor radio access networks," in *Proc. IEEE VTC Spring*, 2015.
- [5] K. Niu, K. Chen, J. Lin, and Q. T. Zhang, "Polar codes: Primary concepts and practical decoding algorithms," *IEEE Commun. Mag.*, vol. 52, no. 7, pp. 192–203, Jul. 2014.
- [6] G. Wu, C. Yang, S. Li, and G. Li, "Recent advance in energy-efficient networks and its application in 5G systems," *IEEE Wireless Commun. Mag.*, vol. 22, no. 2, pp. 145 – 151, Apr. 2015.
- [7] L. Liu, Y. Yi, C. J.-F., and J. Zhang, "Energy-efficient power allocation for delay-sensitive multimedia traffic over wireless systems," *IEEE Trans. Veh. Technol.*, vol. 63, no. 5, pp. 2038 – 2047, Mar. 2014.
- [8] C. She, C. Yang, and L. Liu, "Energy-efficient resource allocation for MIMO-OFDM systems serving random sources with statistical QoS requirement," *IEEE Trans. Commun.*, vol. 63, no. 11, pp. 4125–4141, Nov. 2015.
- [9] C. Chang and J. A. Thomas, "Effective bandwidth in high-speed digital networks," *IEEE J. Sel. Areas Commun.*, vol. 13, no. 6, pp. 1091–1100, Aug. 1995.
- [10] G. L. Choudhury, D. M. Lucantoni, and W. Whitt, "Squeezing the most out of ATM," *IEEE Trans. Commun.*, vol. 44, no. 2, pp. 203–217, Feb. 1996.
- [11] M. Khabazian, S. Aissa, and M. Mehmet-Ali, "Performance modeling of safety messages broadcast in vehicular ad hoc networks," *IEEE Trans. Intell. Transp. Syst.*, vol. 14, no. 1, pp. 380 – 387, Mar. 2013.
- [12] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Channel coding rate in the finite blocklength regime," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2307–2359, May 2010.
- [13] D. Wu and R. Negi, "Effective capacity: A wireless link model for support of quality of service," *IEEE Trans. Wireless Commun.*, vol. 2, no. 4, pp. 630–643, Jul. 2003.
- [14] D. Gross and C. Harris, *Fundamentals of Queueing Theory*. Wiley, 1985.
- [15] G. Auer, O. Blume, V. Giannini, I. Gódor, *et al.*, "D 2.3: Energy efficiency analysis of the reference systems, areas of improvements and target breakdown," *EARTH*, Jan. 2012. [Online]. Available: <https://www.ict-earth.eu/publications/deliverables/deliverables.html>
- [16] B. Debaillie, C. Desset, and F. Louagie, "A flexible and future-proof power model for cellular base stations," in *Proc. IEEE VTC Spring*, 2015.
- [17] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge Univ. Press, 2004.
- [18] F. Rusek, D. Persson, B. K. Lau, E. G. Larsson, T. L. Marzetta, O. Edfors, and F. Tufvesson, "Scaling up MIMO: Opportunities and challenges with very large arrays," *IEEE Signal Process. Mag.*, vol. 30, no. 1, pp. 40 – 60, Jan. 2013.
- [19] C. She and C. Yang, "Optimal EE-delay relation in wireless systems," in *Proc. IEEE Online GreenComm*, Nov. 2015.
- [20] G. Zhang, T. Q. S. Quek, A. Huang, M. Kountouris, and H. Shan, "Delay modeling for heterogeneous backhaul technologies," in *Proc. IEEE VTC Fall*, 2015.